LLM Evaluators Recognize and Favor Their Own Generations

Year: 2024 | Citations: 330 | Authors: Arjun Panickssery, Samuel R. Bowman, Shi Feng

Abstract

Self-evaluation using large language models (LLMs) has proven valuable not only in benchmarking but also methods like reward modeling, constitutional AI, and self-refinement. But new biases are introduced due to the same LLM acting as both the evaluator and the evaluatee. One such bias is self-preference, where an LLM evaluator scores its own outputs higher than others' while human annotators consider them of equal quality. But do LLMs actually recognize their own outputs when they give those texts higher scores, or is it just a coincidence? In this paper, we investigate if self-recognition capability contributes to self-preference. We discover that, out of the box, LLMs such as GPT-4 and Llama 2 have non-trivial accuracy at distinguishing themselves from other LLMs and humans. By fine-tuning LLMs, we discover a linear correlation between self-recognition capability and the strength of self-preference bias; using controlled experiments, we show that the causal explanation resists straightforward confounders. We discuss how self-recognition can interfere with unbiased evaluations and AI safety more generally.